On Discriminative and Semi-Supervised Dimensionality Reduction

Chris Pal, Michael Kelm, Xuerui Wang, Greg Druck and Andrew McCallum
Department of Computer Science,
University of Massachusetts, Amherst, MA 01003

Abstract

We are interested in using the goal of making predictions to influence dimensionality reduction procedures. A number of new methods are emerging aimed at combining attributes of generative and discriminative approaches to data modeling. New approaches to semi-supervised learning have also been emerging. We present and apply some new methods to non-linear and richly structured problems comparing and contrasting models designed for computer vision with those designed for text processing and discuss essential properties that need to be preserved when reducing dimensionality.

Overview

Recently there has been a flurry of interest in exploring new techniques combining generative and discriminative methods through novel model structures and objective functions [6, 1, 7]. As well, new and related semi-supervised methods are emerging such as: entropy regularization [4], which aims to avoid violations of clustering assumptions, information regularization [2], which aims to put decision boundaries in low density areas and [13], an approach based on graph Laplacian methods which aims to achieve label smoothness. We focus here on obtaining dimensionality reductions that are altered by labeled data and which improve classification performance in the context of computer vision and text processing applications.

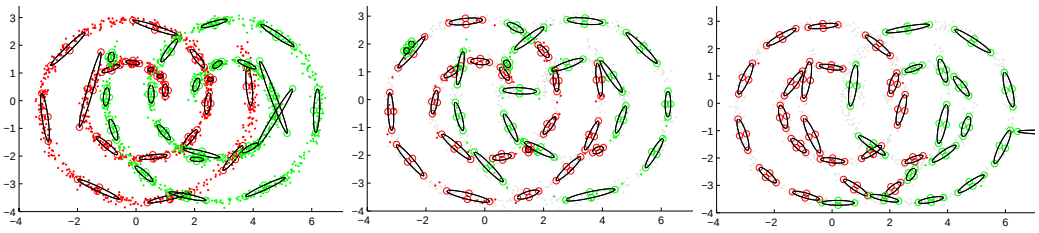

Figure 1: Interleaved Spirals and Dimensionality Reduction. (Left) An MCL based locally linear dimensionality reduction (LLDR). (Middle) Semi-supervised LLDR where grey points illustrate un-labeled data – Here 90% of labels are unobserved. (Right) Here 99% of labels are unobserved. Please note: figure best viewed in color as circles at ellipse extremes indicates subspace class membership and a 3D example is also available.

We begin with an intuitive two class interleaved spiral example, a surrogate problem with structure similar to a variety of computer vision tasks ranging from pixel classification to face manifolds[10]. For example, in [11] mixtures of locally linear models for dimensionality reduction were used for image compression. In the experiments of figure 1 we used a mixture of locally linear, factor analysis models [3] with a single latent dimension or factor, indicated by the dominant axis of the ellipse. The joint distribution for a mixture of factor analyzers can be written as $p(\mathbf{x}, \mathbf{z}, \mathbf{c}, \mathbf{s}) = \exp\{\boldsymbol{\theta}^T \mathbf{c} +$

$\mathbf{c}^T\mathbf{\Theta}\mathbf{s}\}\mathcal{N}(\mathbf{x}, \boldsymbol{\mu}_s + \mathbf{\Lambda}_s\mathbf{z}, \mathbf{\Psi})\mathcal{N}(\mathbf{z}, 0, \mathbf{I})$, with cluster mean $\boldsymbol{\mu}_s$, factor matrix $\mathbf{\Lambda}_s$, latent subspace $\mathbf{z}$, indexed by $\mathbf{s}$, diagonal covariance matrix $\mathbf{\Psi}$ and identity latent space covariance $\mathbf{I}$. The model is an exponential family mixture and can be illustrated with the factor graph shown in fig. 2 (Left). We use these models partly because of their similarities to other locally linear techniques such as locally linear embedding [9]. However, our probabilistic formulation allows us to optimize the model using multi-conditional learning (MCL) [7] – under an objective based on $\alpha \log P(\mathbf{c}|\mathbf{x}) + \beta \log P(\mathbf{x}|\mathbf{c})$, where $\mathbf{x}$ is a continuous input vector, $\mathbf{c}$ is a multinomial class label and α and β are weights selected by hand or using cross-validation methods, typically leading to $\alpha > \beta$. Fig. 1 (Left) illustrates a model obtained using $\alpha = 1, \beta = .05$. Fig. 1 illustrates how most of the models power is devoted to creating a high fidelity decision boundary, with more components in boundary regions. The marginal density of $\mathbf{x}$ is more coarsely captured. To use this underlying model for semi-supervised learning we have experimented with the objective $\alpha \log P(\mathbf{c}|\mathbf{x}) + \beta \log P(\mathbf{x})$. For both we use expected gradient based optimization. In fig. 1 (Middle) and (Right) we obtained models using this objective with $\alpha = 1, \beta = .1$ and with 90% and 99% unlabeled data. Under this metric, even with 90% of the labels missing, the model does a good job at recovering complex non-linear latent spaces and discriminative boundaries. Our quantitative experiments also confirm that both these objectives produce models with superior classification performance compared to Maximum Likelihood. Using this approach, we therefore achieve locally linear but globally non-linear *discriminative* dimensionality reductions using labels to directly improve our model of $p(\mathbf{c}|\mathbf{x})$. Next, we explore a dimensionality reduction for text where cosine comparisons in the latent space must be meaningful for predictions.

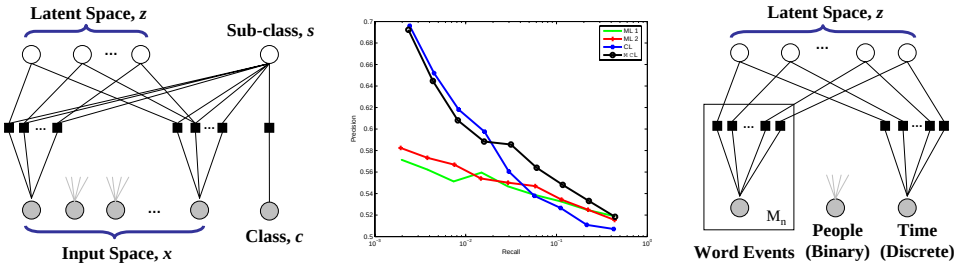

Figure 2: (Left) A graphical model for a mixture of linear subspaces (Middle) Classifying the year of a NIPS paper using a latent space found under CL, MCL, and ML optimization. We see that CL and MCL have clear and superior performance to traditional ML objectives. (Right) The model as a Factor Graph.

Exponential family models [12] and factor graphs [5] are extremely flexible and allow us to create a low dimensional, continuous latent space $\mathbf{z}$ for a variety of richly structured inputs. Consider now creating a latent space that views documents and authors (from NIPS papers) as input and helps us predict publication time (volume number). Consider an input space consisting of a multivariate Bernoulli (binary) random variable $\mathbf{x}_b$ for author identities, a single discrete label $\mathbf{x}_d$ for time and M_n draws from a discrete or multinomial random variables for words. If we define the sum of the word multinomials as $\mathbf{x}_m$, the complete composite input space can be written $\mathbf{x}^T = [\mathbf{x}_b^T\ \mathbf{x}_m^T\ \mathbf{x}_d^T]^T$. If we integrate out the latent space $\mathbf{z}$, we can then write the probability model of fig. 2 (Right) as $P(\mathbf{x}|\boldsymbol{\theta}, \mathbf{\Lambda}) = \exp\{\boldsymbol{\theta}^T\mathbf{x} + \mathbf{x}^T\mathbf{\Lambda}\mathbf{x} - A(\boldsymbol{\theta}, \mathbf{\Lambda})\}$, where $\mathbf{\Lambda} = \frac{1}{2}\mathbf{W}\mathbf{W}^T$, $\mathbf{W}^T = [\mathbf{W}_b^T\mathbf{W}_m^T\mathbf{W}_d^T]$ and $A(\boldsymbol{\theta}, \mathbf{\Lambda})$ is the partition function. To find $\mathbf{W}$, we use Gibbs sampling in an approximate expected gradient based optimization with momentum terms and annealing to speed up convergence. We test our model on the NIPS Conference Papers data set from Roweis [8]. We processed this data set so that: 1) only authors who published (by themselves or co-authored with someone else) more than 5 NIPS papers are used; giving us 125 authors, 2) only papers authored by one or more of the 125 authors are considered, leading to 873 papers, then 3) we select the top 150 words in terms of mutual information for authors. Papers are labeled by the NIPS volume number in which they were published. We retrieve documents that have the same volume label as a test document based on the cosine coefficient between them in the latent space. For evaluation, we score papers as relevant if they are within ± 3 years of when the test document was published. We show precision and recall results in fig. 2. We experiment with Conditional Likelihood (CL) based optimization, using the probability of the volume given authors and words and MCL, where we also use the reverse with $\alpha = 1, \beta = .001$. ML-1 is Maximum likelihood optimization with no volume label, ML-2 uses the volume label. We find that CL and MCL derived latent spaces show marked improvement.

Acknowledgements

This work was supported in part by the Center for Intelligent Information Retrieval, in part by The Central Intelligence Agency, the National Security Agency and National Science Foundation under NSF grant #IIS-0326249, and in part by the Defense Advanced Research Projects Agency (DARPA), through the Department of the Interior, NBC, Acquisition Services Division, under contract number NBCHD030010. This work is also supported in part by Microsoft Research under the eScience and Memex funding programs and by Kodak. Any opinions, findings and conclusions or recommendations expressed in this material are the author(s) and do not necessarily reflect those of the sponsor.

References

[1] G. Bouchard and B. Triggs. The tradeoff between generative and discriminative classifiers. In J. Antoch, editor, *Proceedings in Computational Statistics, 16th Symposium of IASC*, volume 16, Prague, 2004. Physica-Verlag.

[2] A. Corduneanu and T. Jaakkola. On information regularization. In *Proceedings of the 19th UAI, 2003.*, 2003.

[3] Z. Ghahramani and G. E. Hinton. The EM algorithm for mixtures of factor analyzers. Technical Report CRG-TR-96-1, University of Toronto, 1996.

[4] Y. Grandvalet and Y. Bengio. Semi-supervised learning by entropy minimization. In *Advances in Neural Information Processing Systems 17 [Neural Information Processing Systems, NIPS 2004, December 13-18, 2004, Vancouver, British Columbia, Canada]*, 2004.

[5] F. R. Kschischang, B. Frey, and H. A. Loeliger. Factor graphs and the sum-product algorithm. *IEEE Trans. Inform. Theory*, 47(2):498–519, 2001.

[6] J. A. Lasserre, C. M. Bishop, and T. P. Minka. Principled hybrids of generative and discriminative models. In *CVPR '06: Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 87–94, Washington, DC, USA, 2006. IEEE Computer Society.

[7] A. McCallum, C. Pal, G. Druck, and X. Wang. Multi-conditional learning: Generative/discriminative training for clustering and classification. In *To appear in AAAI '06: American Association for Artificial Intelligence National Conference on Artificial Intelligence*, 2006.

[8] http://www.cs.toronto.edu/ roweis/data.html.

[9] S. Roweis and L. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, Dec. 22 2000.

[10] H. S. Seung and D. D. Lee. The manifold ways of perception. *Science*, 290(5500):2268–2269, Dec. 2000.

[11] M. E. Tipping and C. M. Bishop. Mixtures of probabilistic principal component analysers. *Neural Computation*, 11(2):443–482, 1999.

[12] M. Welling, M. Rosen-Zvi, and G. Hinton. Exponential family harmoniums with an application to information retrieval. In *NIPS17*, pages 1481–1488. 2005.

[13] X. Zhu and J. Lafferty. Harmonic mixtures: combining mixture models and graph-based methods for inductive and scalable semi-supervised learning. In *ICML '05: Proceedings of the 22nd international conference on Machine learning*, pages 1052–1059, New York, NY, USA, 2005. ACM Press.